국립국어원 공고 제2019-167호

국립국어원 웹 말뭉치 구축 사업 관련 안내

국립국어원에서는 우리말 인공지능 기술 개발과 국어 연구 등에 활용하고자 대규모 웹 언어 자료를 수집하여 가공하는 '웹 말뭉치 구축' 사업을 추진하고 있습니다. 우리말 인공지능 기술 발전 등에 기초가 되는 국가적 언어 자료의 수집과 구축 사업에 귀하의 소중한 저작물이 유용하게 활용될 수 있도록 많은 관심과 참여를 부탁드립니다.

□ 사업 개요

○ 사업명: 웹 말뭉치 구축

○ 사업 기간: 2019. 5. 27. ~ 2019. 11. 23.

○ 사업 수행자: ㈜메트릭스코퍼레이션

○ 주요 사업 내용

- 웹 원문 자료 수집
  - · 누리소통망(SNS) 200만 게시물 수집
  - · 블로그 1만 게시물 수집
  - · 게시판 1판 게시물 수집
  - · 상품평 등 리뷰 10만 게시물 수집
- 수집된 웹 언어 자료를 대상으로 기초(원시) 말뭉치* 구축

* **말뭉치**: 컴퓨터가 읽을 수 있는 형태로 입력, 분석한 대규모 언어 자료로, 사람에게 학습용 책이 필요하듯 인공지능은 학습용 대규모 언어 자료가 필요함.

○ 담당자: 국립국어원 언어정보과 홍혜진(02-2669-9756)

□ 주요 질의 · 답변

1. 웹 언어 자료를 수집하는 목적은?

○ 누리소통망(SNS), 블로그, 게시판 등에서 실제로 사용된 웹 언어 자료를 모아 컴퓨터가 읽을 수 있는 형태로 분석한 **말뭉치**를 국가적으로 구축하여 **우리말 인공지능 개발과 국어 연구 등에 공공 자료로 활용**할 수 있도록 하기 위해서입니다.

2. 저작권 이용 허락 범위는?

○ **국립국어원과 국립국어원의 용역 사업 수행자**가 귀하의 웹 언어 자료를 말뭉치로 구축하고 배포하기 위하여 아래 일을 할 수 있도록 허락을 해 주시는 것이 필요합니다.

- 수집 자료를 일정한 형식으로 **전자적 기록 매체에 담아 보존**하는 일
- 수집 자료를 형태소, 단어, 문장 등의 **언어 단위별로 분리**하며, **언어적·비언어적 정보를 부착**하는 등 자료를 복제하여 변형하여 **말뭉치를 구축**하는 일
- 구축된 **말뭉치를 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포**하는 일

○ **학계·연구기관·산업체** 등이 국어 연구와 언어 정보 처리 분야 응용 등을 위하여 아래 일을 할 수 있도록 허락을 해 주시는 것이 필요합니다.

- 우리말 인공지능 기술 개발과 국어 연구용으로 **말뭉치를 분석 및 처리하여 사용**하도록 하는 일

3. 저작권 이용 허락 기간은?

○ 학계·연구기관·산업체 등이 연구 및 기술 개발에 활용하기 위해서는 **충분한 기간 동안 안정적으로 말뭉치를 이용할 수 있는 것이 중요**합니다. 예를 들어 1990년대 초반에 영국에서 구축한 BNC(British National Corpus) 말뭉치는 25년이 지난 현재까지도 안정적으로

제공되어 활용되고 있습니다. 국립국어원에서는 귀하의 소중한 웹 언어 자료를 말뭉치로 구축하여 최소 **2035년 12월 31일까지는 안정적으로 이용**할 수 있도록 허락해 주시기를 바랍니다.

○ 귀하께서 **이용 허락 중지 의사를 밝히시면 최소 이용 허락 기간이 끝난 후 즉시 이용을 중지**할 예정입니다.

4. 웹 말뭉치는 어떠한 형식으로 구축되는 것인지?

○ 귀하께서 작성하신 웹 언어 자료의 **원문을 수집**하고, 수집된 자료에 **말뭉치의 형식을 갖추기 위한 정보를 부가**하여 원시 말뭉치로 구축합니다. 여기에 형태소, 어휘, 문장과 관련된 **언어적 정보를 부가**하여 분석 말뭉치로 구축할 수 있습니다.

<원시 말뭉치 예시>

```
<?xml version="1.0" encoding="UTF-8"?>
<SJML>
  <header>
    <fileInfo>
      <fileId>ESRW1900000001</fileId>
      <annoLevel>원시</annoLevel>
      <class>누리소통망</class>
    </fileInfo>
    <sourceInfo>
      <title>오늘은 이상하게</title>
      <author>Hyejin Hong</author>
      <publisher>페이스북</publisher>
      <date>2013. 03. 08. 11:41</date>
      <dateCrawl>2019. 07. 16. 09:36</dateCrawl>
      <view>4</view>
    </sourceInfo>
  </header>
  <text>
    <p>오늘은 이상하게 바닥에 떨어져있는 단추가 많네. 다들 오래 넣어둔 봄옷을 꺼내 입으셔서 그런가.</p>
  </text>
</SJML>
```

5. 개인정보가 노출될 우려는 없는지?

○ 이름, 전화번호, 주소 등 **개인정보는 철저하게 알아볼 수 없게 처리합니다.**